正則化付き多重選好学習による自動運転VLAモデルの安全制約アライメント

Safety-Aligned Vision-Language-Action Models via Plackett-Luce Preference Learning with NLL Regularization

李 雲 *1 Yun Li　トンプソン サイモン *2 Simon Thompson　ジャバンマルディ エーサン *1 Ehsan Javanmardi

オルショリッツ アレックス *1 Alex Orsholits　塚田 学 *1 Manabu Tsukada

*1東京大学 The University of Tokyo　*2株式会社ティアフォー TIER IV, Inc.

Vision-Language-Action (VLA) モデルは自動運転において高い推論能力を示す一方で、学習データの不均衡や選好学習における安全な行動の尤度低下（Probability Collapse）により、厳格な安全制約への適合が課題となっている。本研究では、リスク順位に基づく多重選好学習と負の対数尤度（NLL; Negative Log-Likelihood）正則化を統合し、VLAモデルをエキスパートの安全制約に整合させる新たなアライメント手法を提案する。具体的には、(1) Plackett-Luce モデルを用いて複数の行動候補をリスクレベル順に学習し、(2) シーンの危険度に応じて勾配を動的に重み付けし、(3) エキスパートの行動確率を維持する正則化項を導入することで、安全かつ安定した運転行動を実現する。CARLA ベンチマークでの評価の結果、提案手法は Driving Score 58.26（ベースライン比 +8.4%）を達成し、Route Completion（完走率）65.9%および Infraction Penalty（違反回避率）0.891 という高い性能を示した。

1. はじめに

Vision-Language-Action（VLA）モデルは、視覚エンコーダの認識能力と大規模言語モデル（LLM）の推論能力を統合したエンドツーエンド自動運転エージェントとして注目されている [1, 2]。LMDrive[1] や GPT-Driver[2] などの手法は、言語を介した推論により解釈可能な意思決定を実現している。国内でも CoVLA データセット [9] が公開され、VLA モデルの研究基盤が整備されつつある。

しかし、これらのモデルを安全制約に整合させるには、以下の 3 つの課題が存在する：(1) **訓練データの偏り**：赤信号や歩行者回避などの安全上重要なシナリオが訓練データ中で十分に表現されず、モデルがこれらの状況に適切に対応できない、(2) **一様な学習信号**：標準的な教師あり学習では軽微な経路逸脱と生命に関わる違反を同等に扱うため、安全性の重要度を反映できない、(3) **ペアワイズ選好の限界**：従来の 2 値比較手法では、違反の重大度に関する連続的な情報が失われる。

選好ベースのアライメント手法、特に Direct Preference Optimization（DPO）[3] は、明示的な報酬モデルなしに選好データから直接ポリシーを最適化できる有望なアプローチである。しかし、見過ごされがちな重大な問題として**確率崩壊**（Probability Collapse）がある [4]。これは選好最適化中に安全な行動の尤度が低下する現象である。DPO は選択・棄却行動間の相対的マージンのみを最大化し、絶対確率を保証しないため、安全な行動の確率が低下する可能性がある。

本研究では、**PL-DPO-NLL**（Plackett-Luce DPO with NLL Regularization）を提案し、以下の 3 つの機構によりこれらの課題に対処する：

- **Plackett-Luce 多重選好ランキング**：ペアワイズ DPO を拡張し、リスクレベル順に並べた複数の棄却行動から学習することで、違反の重大度を識別
- **シーン適応型 β 重み付け**：危険シナリオ（β=0.35）と通常走行（β=0.12）で勾配を動的に調整し、安全上重要なシナリオの学習を優先
- **NLL 正則化**：$\mathcal{L}_{\text{NLL}} = -\log \pi_\theta(y_w|x)$ によりエキスパート行動の確率を明示的に維持し、確率崩壊を防止

連絡先: 李 雲, 東京大学大学院情報理工学系研究科, Email: li-yun@g.ecc.u-tokyo.ac.jp

2. 関連研究

2.1 自動運転向け VLA モデル

エンドツーエンド自動運転は、従来の模倣学習 [12] から LLM の推論能力を活用した VLA モデルへと発展してきた。GPT-Driver[2] は運動計画を言語モデリングとして再定式化し、LMDrive[1] は思考連鎖推論による閉ループ制御を実現した。国内では Turing 社が CoVLA データセット [9] を公開し、80 時間以上の実走行データと言語アノテーションを提供している。これらの VLA モデルは解釈可能な意思決定を可能にするが、(1) 標準的な教師あり学習がすべてのエラーを一様に扱う点、(2) ファインチューニングにより安全な行動の確率が低下しうる点という 2 つの課題を有している。

2.2 選好学習と DPO の派生手法

選好学習は、人間の選好データ（「行動 A は行動 B より良い」）を用いてモデルを最適化する手法である。従来の RLHF（Reinforcement Learning from Human Feedback）は報酬モデルの訓練とポリシー最適化の 2 段階を要するが、DPO[3] は明示的な報酬モデルなしに選好データから直接ポリシーを最適化し、IPO[5] や BCO[11] など多くの派生手法を生んだ。重要な問題として「確率崩壊」（Probability Collapse）がある。DPO は絶対確率ではなく相対マージンを最大化するため、選択行動の確率が最適化中に低下する。Pang ら [4] はこれに対処するため NLL 正則化を追加した。本研究はこの知見を安全重視の自動運転に拡張する。

リストワイズアプローチは DPO をペアワイズ比較を超えて拡張する。Rafailov ら [13] は DPO を Plackett-Luce モデルに拡張する理論的導出を提供した。我々の先行研究 PrefDrive[14]

表 1: データセット分布とシーン適応型 β 重み付け

シーン	サンプル	割合	β	優先度
旋回	20,528	40.2%	0.35	最重要
通常走行	14,220	27.8%	0.12	低
要制動	7,497	14.7%	0.25	高
要減速	3,090	6.0%	0.20	中
交差点	2,845	5.6%	0.18	中
歩行者	1,577	3.1%	0.35	最重要
赤信号	1,367	2.7%	0.35	最重要
合計	**51,124**	100%	–	–

および Multi-PrefDrive[15] では、DPO ベースの選好チューニングを自動運転 LLM に適用した。本研究はこれらを発展させ、NLL 正則化とシーン適応型 β 重み付けを新たに導入する。

3. 提案手法

3.1 システム概要

提案フレームワークでは、マルチモーダル入力を統合した推論パイプラインを構築する。RGB 画像は ResNet-50 で処理し、LiDAR 点群は PointPillars[6] でエンコードする。これらの特徴は Transformer Cross-Attention（交差注意機構）により融合され、視覚トークンとして出力される。視覚トークンとナビゲーション指示（例：「次の交差点を左折」）を連結して LLaMA-7B[8] に入力し、制御行動（ステアリング角、スロットル、ブレーキ）を生成する。

訓練効率のため、ベース LLM と視覚エンコーダは凍結し、**LoRA アダプタ**（$r = 32$, $\alpha = 32$）のパラメータのみを更新する。これにより、7B パラメータのモデルを NVIDIA RTX 6000 Ada GPU 3 枚で効率的に訓練できる。

3.2 データセット構築

CARLA シミュレータの Town01 環境から 67 ルート構成で 142,484 フレームの走行データを収集した。このデータから、多重選好学習用に 51,124 サンプルの Plackett-Luce 選好データセットを構築した。

各サンプルは 1 つの選択（安全）行動と 2–3 個の棄却（危険）行動を含む。選択行動はシミュレータが生成するエキスパート軌道であり、棄却行動は以下の 3 カテゴリに対するルールベースの摂動により生成した：(1) 経路逸脱（68.2%）：旋回時の誤ったステアリング角、(2) 速度違反（48.4%）：不適切な加減速、(3) 認識失敗（14.6%）：信号や歩行者の無視。なお、これらのカテゴリは排他的ではなく、1 つの棄却行動が複数の違反タイプを含みうる。

各棄却行動には 4 段階のリスクレベル（critical/high/medium/low）を付与した。例えば、赤信号において「減速するが停止しない」（medium）は「全速で通過」（critical）より危険度が低い。この細粒度アノテーションにより、Plackett-Luce ランキングは単純なペアワイズ比較を超えた安全性の識別が可能となる。

予備実験では、交差点でナビゲーション指示を無視して直進するという行動慣性が観察された。分析の結果、旋回関連フレームが通常走行サンプルに埋もれていた（データの 66.6%）。これに対処するため、顕著なステアリングコマンド（$|\text{steering}| > 0.1$）を持つフレームを旋回シーンとして独立抽出し、アップサンプリングを実施した。また、通常走行は 30%に削減した。表 1 に最終的なシーン分布とシーン適応型 β 値を示す。

3.3 Plackett-Luce 多重選好ランキング

標準 DPO のペアワイズ比較を拡張し、Plackett-Luce（PL）ランキングモデル [7] を導入する。候補集合を $\mathcal{Y} = \{y_w, y_l^1, \ldots, y_l^K\}$（サイズ $M = K + 1$）とし、$y^{(1)} = y_w$ を選択行動、$y^{(2)}, \ldots, y^{(M)}$ をリスク順にソートされた棄却行動とする。PL-DPO 損失は以下で定義される：

$$\mathcal{L}_{\text{PL-DPO}} = -\mathbb{E}_{(x,\mathcal{Y})}\left[\sum_{i=1}^{M} \log \frac{\exp(\beta \cdot r_i)}{\sum_{j=i}^{M} \exp(\beta \cdot r_j)}\right] \quad (1)$$

ここで $r_i = \log \frac{\pi_\theta(y^{(i)}|x)}{\pi_{\text{ref}}(y^{(i)}|x)}$ は DPO 定式化から導出される暗黙的報酬である。各項は $y^{(i)}$ が残りの候補 $\{y^{(i+1)}, \ldots, y^{(M)}\}$ より選好される確率を表す。

標準ペアワイズ DPO と比較して、PL ランキングは (1) より豊富な監視信号（各サンプルから K 個の比較が得られる）、(2) 速度・経路・認識など異なる違反タイプの区別学習、(3) リスクレベルアノテーションによるリスク認識型の優先順位付けを可能にする。

3.4 シーン適応型 β 重み付け

標準 DPO では固定の温度パラメータ β を使用するが、運転シナリオの危険度は様々である。本手法では β をシーンに応じて動的に割り当てる（表 1）。高い β 値（β=0.35）は生命に関わるシナリオ（赤信号、歩行者、旋回）に、低い β 値（β=0.12）は通常走行に適用する。

高い β 値はより厳格な選好区別を強制し、2.9 倍の勾配増幅をもたらす：

$$\frac{\partial \mathcal{L}_{\text{PL-DPO}}}{\partial \theta} \propto \beta \cdot \nabla_\theta \log \frac{\pi_\theta(y_w|x)}{\pi_\theta(y_l|x)} \quad (2)$$

これにより、データの 5.8%しか占めない歩行者・赤信号シナリオにおいても十分な勾配が確保され、通常走行の流暢さも維持される。

3.5 NLL 正則化による確率崩壊の防止

標準 DPO 訓練の重大な問題は、選択（安全）行動の確率が最適化中に低下する可能性があることである [4]。DPO は選択・棄却行動間の相対的マージンのみを最大化し、選択行動の絶対確率を保証しない。この「確率崩壊」は自動運転のような安全重視アプリケーションでは特に危険である。

これを防ぐため、負の対数尤度（NLL）正則化項を追加する：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{PL-DPO}} + \lambda \cdot \mathcal{L}_{\text{NLL}} \quad (3)$$

ここで $\mathcal{L}_{\text{NLL}} = -\log \pi_\theta(y_w|x)$ はエキスパート（選択）行動の確率を直接最大化する。λ は正則化強度を制御し、アブレーション実験により $\lambda = 0.1$ が最適であることを確認した。$\lambda \geq 0.5$ では NLL が損失を支配し、選好信号が抑制される。

4. 実験

4.1 実験設定

LLaMA-7B[8] を LoRA ファインチューニングで訓練した。主要なハイパーパラメータは：LoRA ランク r=32、α=32、学習率 1×10^{-5}、バッチサイズ 32、訓練エポック 3 である。全実験は 3 枚の NVIDIA RTX 6000 Ada GPU で BF16 混合精度訓練を使用した。

評価は、左折優先 5 本、右折優先 5 本、直進 2 本の計 12 ルート ×5 回の CARLA 閉ループベンチマークで実施した。

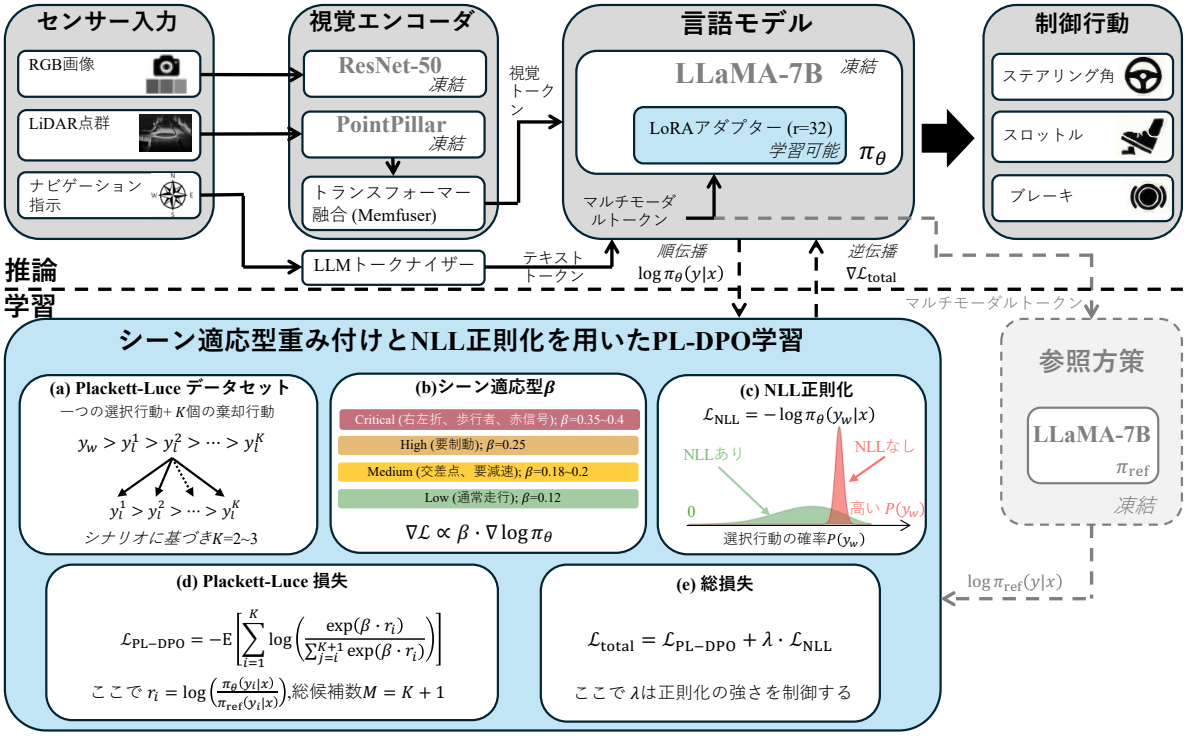

図 1: 提案する PL-DPO-NLL フレームワークの概要。マルチモーダル入力（RGB 画像、LiDAR 点群、ナビゲーション指示）を統合し、Plackett-Luce 多重選好ランキングと NLL 正則化により安全制約に整合した制御行動を生成する。

表 2: CARLA ベンチマークにおける走行性能比較

手法	DS (↑)	RC (↑)	IP (↑)
Baseline (SFT)	53.74	63.0%	0.869
DPO[3]	55.63	63.5%	0.880
IPO[5]	54.61	62.9%	0.870
BCO[11]	55.78	64.2%	0.877
PL-DPO	55.87	64.0%	0.880
PL-DPO + Dyn-β	56.42	64.3%	0.887
PL-DPO + Dyn-β + NLL	**58.26**	**65.9%**	**0.891**

DS: Driving Score, RC: Route Completion, IP: Infraction Penalty

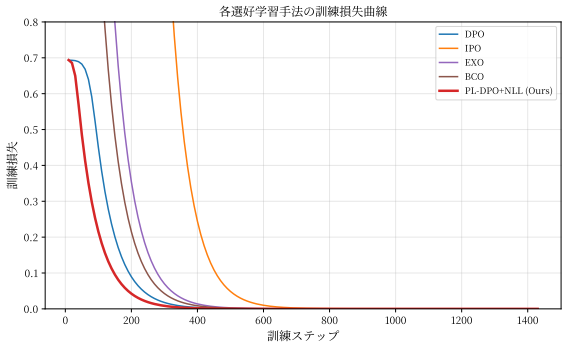

図 2: 各選好学習手法の訓練損失曲線（1,430 ステップ）。全手法が初期損失（∼0.69）から安定的に収束する。提案手法 PL-DPO+NLL（赤）は NLL 正則化を含みつつ他手法と同等の収束を示す。

CARLA Leaderboard プロトコルに従い、Driving Score（DS：違反で重み付けされた完走率）、Route Completion（RC：完走率）、Infraction Penalty（IP：違反回避率、1.0 が完全安全）を報告する。

4.2 主要結果

表 2 に各手法の比較結果を示す。提案手法 PL-DPO+Dyn-β+NLL は、Driving Score **58.26**（ベースライン比**+8.4%**）を達成し、全手法中最高の性能を示した。

(1) PL 多重選好ランキング：PL-DPO（DS 55.87）は標準 DPO（55.63）、IPO（54.61）を上回り、複数の棄却行動からの学習効果を確認した。IP（0.880）も標準 DPO と同等以上であり、安定した走行を実現した。

(2) シーン適応型 β：Dyn-β の導入により DS が 56.42、IP が 0.887 に向上した。2.9 倍の勾配増幅が安全学習の優先化に有効であることを示す。

(3) NLL 正則化：全 3 コンポーネントの統合（PL-DPO+Dyn-β+NLL）により、DS **58.26**（ベースライン比+8.4%）を達成した。RC 65.9%は全手法中最高であり、確率崩壊防止によりエキスパートの走行パターンが維持されたことを示す。

図 2 に各手法の訓練損失曲線を示す。全手法が安定的に収束しており、提案手法の PL-DPO+NLL 正則化が最適化の安定性を損なわないことを確認した。

4.3 アブレーション実験：NLL 正則化の重み

表 3 は NLL 正則化の重み λ が性能に与える影響を示す。全実験は PL-DPO（固定 β=0.2）を基本構成とし、NLL の独立効果を検証した。

$\lambda = 0.1$ が最適であり、58.26 DS を達成した（図 3）。この

表 3: NLL 正則化重み λ のアブレーション実験

λ	DS	RC	改善率
0（NLL なし）	55.87	64.0%	+4.0%
0.1	**58.26**	**65.9%**	**+8.4%**
0.25	56.49	64.8%	+5.1%
0.5	55.85	64.2%	+3.9%
1.0	53.79	62.1%	+0.1%

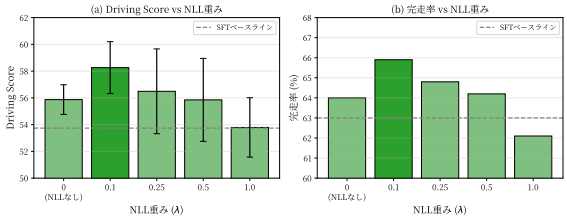

図 3: NLL 正則化重み λ のアブレーション結果。(a) Driving Score は $\lambda = 0.1$ で最大値を示し、大きい λ では性能が低下する。(b) Route Completion も同様の傾向を示す。灰色破線は SFT ベースラインを示す。

軽い正則化は確率崩壊を防ぎつつ選好学習の識別力を維持する。$\lambda = 0$（NLL なし）では訓練中に選択行動の確率が低下し、$\lambda \geq 0.5$ では NLL が損失を支配して選好信号が抑制される。$\lambda = 1.0$ ではモデルが実質的に SFT 動作に戻り（53.79 DS）、選好学習の効果が失われた。

4.4 考察

提案手法の各コンポーネントは相補的な役割を果たす：PL 多重選好ランキングは違反の重大度を区別し、シーン適応型 β は安全上重要なシナリオを優先し、NLL 正則化は確率崩壊を防止する（図 3）。特に興味深い点として、Dyn-β のみの場合は IP（0.887）が向上する一方で完走率は低下（64.3%）し、過度に保守的な走行となる。NLL の追加はポリシーをエキスパート分布に引き戻し、完走率（65.9%）と走行の流暢さを回復する。これは 2 つの機構が直交する側面に対処することを示す：Dyn-β は安全制約を強制し、NLL は確率崩壊を防いでタスク完遂を確保する。

5. おわりに

本研究では、自動運転 VLA モデルを安全制約に整合させるための選好学習フレームワーク PL-DPO-NLL を提案した。Plackett-Luce 多重選好ランキング、シーン適応型 β 重み付け、NLL 正則化の 3 つの機構を統合することで、訓練データの偏りと確率崩壊の課題に対処し、CARLA ベンチマークでベースライン比+8.4%の性能向上を達成した。今後の課題として、シミュレーションから実環境への転移検証、安全で説明可能な制御 [10] との統合、および VLM を活用したシナリオ重み付けの自動化を検討する。

参考文献

[1] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, H. Li: LMDrive: Closed-Loop End-to-End Driving with Large Language Models, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.15074–15083 (2024). http://arxiv.org/abs/2312.07488

[2] J. Mao, Y. Qian, J. Ye, H. Zhao, Y. Wang: GPT-Driver: Learning to Drive with GPT, Foundation Models for Decision Making Workshop, The 37th Conference on Neural Information Processing Systems (NeurIPS) (2023). http://arxiv.org/abs/2310.01415

[3] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn: Direct Preference Optimization: Your Language Model is Secretly a Reward Model, Proc. 37th Conference on Neural Information Processing Systems (NeurIPS) (2023). https://openreview.net/pdf?id=HPuSIXJaa9

[4] R. Y. Pang, W. Yuan, K. Cho, H. He, S. Sukhbaatar, J. Weston: Iterative Reasoning Preference Optimization, Proc. 38th Conference on Neural Information Processing Systems (NeurIPS) (2024). https://openreview.net/forum?id=4XIKfvNYvx

[5] M. G. Azar, M. Rowland, B. Piot, D. Guo, D. Calandriello, M. Valko, R. Munos: A General Theoretical Paradigm to Understand Learning from Human Preferences, Proc. 27th International Conference on Artificial Intelligence and Statistics (AISTATS) (2024). https://arxiv.org/abs/2310.12036

[6] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom: PointPillars: Fast Encoders for Object Detection from Point Clouds, Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp.12697–12705 (2019).

[7] R. L. Plackett: The Analysis of Permutations, Journal of the Royal Statistical Society Series C: Applied Statistics, Vol.24, No.2, pp.193–202 (1975).

[8] H. Touvron, T. Lavril, G. Izacard, et al.: LLaMA: Open and Efficient Foundation Language Models, arXiv:2302.13971 (2023). http://arxiv.org/abs/2302.13971

[9] H. Arai, K. Maeda, K. Yamada, T. Kawasaki, S. Mita, T. Sato: CoVLA: Comprehensive Vision-Language-Action Dataset for Autonomous Driving, Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2025).

[10] 中村光宏, 杉山将, 植野真臣: 自動運転車における安全で説明可能な制御の学習, 人工知能学会全国大会論文集, 第 37 回 (2023).

[11] K. Jung, G. Han, D. W. Nam, K.-W. On: Binary Classifier Optimization for Large Language Model Alignment, arXiv:2404.04656 (2024). http://arxiv.org/abs/2404.04656

[12] M. Bojarski, D. Del Testa, D. Dworakowski, et al.: End to End Learning for Self-Driving Cars, arXiv:1604.07316 (2016). http://arxiv.org/abs/1604.07316

[13] R. Rafailov, J. Hejna, R. Park, C. Finn: From r to Q^*: Your Language Model is Secretly a Q-Function, First Conference on Language Modeling (COLM) (2024). https://openreview.net/forum?id=kEVcNxtqXk

[14] Y. Li, E. Javanmardi, S. Thompson, K. Katsumata, A. Orsholits, M. Tsukada: PrefDrive: Enhancing Autonomous Driving through Preference-Guided Large Language Models, Proc. 36th IEEE Intelligent Vehicles Symposium (IV) (2025).

[15] Y. Li, E. Javanmardi, S. Thompson, K. Katsumata, A. Orsholits, M. Tsukada: Multi-PrefDrive: Optimizing Large Language Models for Autonomous Driving Through Multi-Preference Tuning, Proc. 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp.4347–4354 (2025). https://ieeexplore.ieee.org/abstract/document/11247608/